Towards the Detection of Diffusion Model Deepfakes

Year: 2022 | Citations: 133 | Authors: Jonas Ricker, Simon Damm, Thorsten Holz, Asja Fischer

Abstract

In the course of the past few years, diffusion models (DMs) have reached an unprecedented level of visual quality. However, relatively little attention has been paid to the detection of DM-generated images, which is critical to prevent adverse impacts on our society. In contrast, generative adversarial networks (GANs), have been extensively studied from a forensic perspective. In this work, we therefore take the natural next step to evaluate whether previous methods can be used to detect images generated by DMs. Our experiments yield two key findings: (1) state-of-the-art GAN detectors are unable to reliably distinguish real from DM-generated images, but (2) re-training them on DM-generated images allows for almost perfect detection, which remarkably even generalizes to GANs. Together with a feature space analysis, our results lead to the hypothesis that DMs produce fewer detectable artifacts and are thus more difficult to detect compared to GANs. One possible reason for this is the absence of grid-like frequency artifacts in DM-generated images, which are a known weakness of GANs. However, we make the interesting observation that diffusion models tend to underestimate high frequencies, which we attribute to the learning objective.